Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Robust Hyperparameter Estimation in Gaussian Process Regression Model

Pooja Algikar (apooja19@vt.edu)
Virginia Tech
Lamine Mili (lmili@vt.edu)
Virginia Tech

Research Article

Keywords:

DOI: https://doi.org/

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Robust Hyperparameter Estimation in Gaussian Process Regression Model

Pooja Algikar[1,1*] and Lamine Mili[1,1]

[1,1]Bradley Department of Electrical and Computer Engineering, Virginia Tech, 7054 Haycock Rd, Falls Church, 22043, VA, USA.

*Corresponding author(s). E-mail(s): apooja19@vt.edu;
Contributing authors: lmili@vt.edu;

Abstract

A stochastic process mathematically models the random behavior of systems and naturally occurring phenomena. A popular process is a Gaussian process which assumes a multivariate Gaussian distribution for the response variables. The predictive distributions of target variables can be explicitly obtained by using conditioning and inference techniques. The mean function and covariance function of the predictive distributions are parameterized whose estimates can be obtained by maximizing the marginal likelihood of the training responses or their posterior distribution. Unfortunately, these estimates are not robust to outliers that are often present in the observational training data sets. To overcome this problem, we develop a robust data-driven process based on the Schweppe-type Huber-generalized maximum likelihood (SHGM) estimator to bound the influence of all types of outliers on the estimates while exhibiting a high statistical efficiency at the Gaussian and thick-tail distributions. The weights utilized by the SHGM estimator are calculated using projection statistics, which are robust distances of the data points determined by the row vectors of the design matrix. These weights make the estimator robust to both vertical outliers and bad leverage points while not downweighting good leverage points. We show that the proposed robust model is able to handle up to 25 percent of outliers among the data set. The proposed method is illustrated by two real-world examples in which the performance of the proposed method is assessed against the standard Gaussian process regression.

1 Introduction

Gaussian processes (GPs) have been increasingly popular over the past few decades for robust regression. A GP model typically consists of a mean function and a kernel (i.e., covariance) function. The parameters of these functions are called hyperparameters. The mean function models the central tendency of the response variable while the kernel function captures the dependencies between the output data points in terms of either input data points or distance between them. Upon estimating the hyperparameters, a GP model is ready to make predictions of response variables for new input data. Numerical and graphical validation techniques have been proposed to assess how well the GP model represents the underlying process (Bastos et al. (2009)).

A crucial factor in determining the performance of a GPs is the quality of the training data set. In the most simplistic form of GPs i.e. the conjugate model, the likelihood of the observations is assumed to be a normal distribution. If the assumed probability distributions do not model well the randomness of the observational data sets of the real process under study, the statistical estimates of the output variables at new test points obtained from the GPs will be inaccurate. For instance, the stochastic power flow studies conducted on an electric power system with high integration of renewable energy resources (Xu et al. (2020)) typically assume the two-parameter Weibull probability distribution for the wind speed and the beta probability distribution for the solar power generation, which are poor statistical models of these variables (Wais (2017)). Such large deviations from the assumptions will result in strongly biased hyperparameter estimates if non-robust estimation methods are applied which eventually affects the resulting uncertainty quantification. Consequently, data-driven models are popular in these areas (see, Wang et al. (2020)).

However, real-world data sets are typically contaminated with outliers due to instrumental failures, human errors, meter sensitivity to changing weather conditions, and recording data errors, to name a few. The two types of outliers likely to occur in a linear regression model of GPs are vertical outliers and bad leverage points. A vertical outlier is an outlier in the response space whose projection on the input vector space associated with the input variables is an inlier. A leverage point is a data point whose projection on the input vector space is an outlier. A leverage point can be a good leverage point or a bad leverage point depending upon whether it is an inlier or an outlier in the response space, respectively. GPs are based on the ℓ_2 norm likelihood functions, which are non-robust to outliers (see, Huber (2004)).

A number of statistical methods are proposed in the literature to robustify the GPs against outliers. For instance, Kuss et al. (2005) proposed GPs with the likelihood of the observed data following Student-t and Laplace distributions. He then estimates the latent parameters of the model using a fully Bayesian treatment. Along similar lines, Jylänki et al. (2011) proposed robust GPs with Student-t likelihood and advanced expectation propagation and Monte Carlo methods algorithms to estimate the latent parameters.

Unfortunately, these methods will breakdown in presence of multiple outliers strategically located in the input vector space as demonstrated by Algikar et al. (2023).

In this article, we stick to the robust type-II maximum likelihood estimation in which the modeling effort is focused on the mean function hyperparameters. We develop a robust observational data-driven process (RODP) regression model that handles vertical outliers and bad leverage points in the training data and for wide-range thick-tailed error distributions. The former is achieved by downweighting the residuals using weights calculated from the projection statistics that are assessed at the hyperparameter estimation step. Introduced by Stahel (1981) and Donoho (1982), a projection statistic of a multivariate data point relative to a multivariate point cloud is defined as the maximum of the absolute values of the standardized projections of that data point on the straight lines that pass through the coordinatewise median and the data points (see, Rousseeuw et al. (2005)). It is a robust version of the Mahalanobis distance based on the sample median and the median absolute deviation from the median instead of the sample mean and the sample standard deviation.

The rest of the article is organized as follows. Section 2 reviews the principal ideas of the GPs and the SHGM estimator. Section 3 develops the RODP model estimator and demonstrates its robustness using an artificial example. Section 4 evaluates the performance of the RODP model on an artificial numerical model and two real-world applications. Section 5 concludes the article and discusses future work. This article has supplementary materials that are available online.

2 BACKGROUND

In this section, we review the GP model and the maximum marginal likelihood estimation, commonly known as the type-II maximum likelihood estimation, of the hyperparameters of the mean and covariance function. And show that the hyperparameters of the mean function are centered around the weighted least squares estimates. We also review the SHGM estimator on which the RODP model is based.

2.1 Maximum Marginal Likelihood Estimation of the Gaussian Process

Let us consider a simulator, $\mathcal{H}(\cdot)$, defined as a function of m input stochastic variables of n dimensions, $\mathbf{x}_i \in \mathbb{R}^n$, and m output stochastic variables, $y_i \in \mathbb{R}$, $i = 1, \ldots, m$. The simulator is statistically approximated by a more computationally efficient model, which may have the same inputs and output variables or functions of these variables.

In the GP framework, the output variables, $\mathcal{H}(\mathbf{x}_i), i = 1, \ldots, m$, are assumed to follow a normal distribution with a vector-valued mean function, $\boldsymbol{m}(\cdot)$, and a matrix-valued covariance function, $\boldsymbol{\Sigma}(\cdot, \cdot)$, for an input variable

matrix, $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_m]^T$, yielding

$$\mathcal{H}(\mathbf{X}) \sim \mathcal{GP}\left(\boldsymbol{m}(\mathbf{X}), \boldsymbol{\Sigma}(\mathbf{X}, \mathbf{X})\right). \tag{1}$$

The elements of the covariance function, $\boldsymbol{\Sigma}$, are kernel functions given by $\Sigma_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j)$, $i, j = 1, 2, \ldots, m$, where $k(\mathbf{x}_i, \mathbf{x}_j)$ models the spatial covariance of the corresponding simulator output points, $\mathcal{H}(\mathbf{x}_i)$ and $\mathcal{H}(\mathbf{x}_j)$. Since the simulator runs, $\mathcal{H}(\mathbf{X})$, are GPs, they follow a joint multivariate Gaussian distribution for all $\mathbf{X}$.

A commonly used squared exponential kernel function is defined as

$$k(\mathbf{x}_i, \mathbf{x_j} \mid \boldsymbol{s}) = \tau^2 R_c(\mathbf{x}_i, \mathbf{x}_j), \tag{2}$$

where $R_c(\mathbf{x}_i, \mathbf{x_j})$ denotes the autocorrelation function given by $R_c(\mathbf{x}_i, \mathbf{x_j}) = \exp\left(-\sum_{k=1}^{p} \frac{(\mathrm{x}_{ik} - \mathrm{x}_{jk})^2}{\boldsymbol{s}_k^2}\right)$. The observations (e.g., measurements), $\mathbf{y}$, of a simulator, $\mathcal{H}(\cdot)$, are assumed to differ from the simulator outputs, $\mathcal{H}(\mathbf{X})$, by additive independent and identically distributed Gaussian errors contained in the vector, $\boldsymbol{e} = [e_1, e_2, \cdots, e_m]^T \sim \mathcal{N}(\mathbf{0}, \sigma_0^2 \mathbf{I}_m)$. Formally, we have

$$\mathbf{y} = \mathcal{H}(\mathbf{X}) + \boldsymbol{e}. \tag{3}$$

"Nuggets" are elements added to the diagonal elements of the covariance function, $\boldsymbol{\Sigma}(\cdot)$, to account for the model uncertainties and the numerical approximations that result from the matrix inverse calculation. They are assumed to follow a Gaussian distribution with zero means, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma_n^2 \mathbf{I}_m)$. Consequently, the observation vector, $\mathbf{y}$, follows a multivariate Gaussian distribution with diagonal additive nugget terms, yielding

$$\mathbf{y} \mid \mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\phi}, \sigma_n^2 \sim \mathcal{N}\left(\boldsymbol{m}(\mathbf{X}), \boldsymbol{\Sigma}(\mathbf{X}, \mathbf{X} \mid \boldsymbol{\phi}) + \sigma_n^2 \mathbf{I}_m\right), \tag{4}$$

where the hyperparameters of the covariance function are grouped into $\boldsymbol{\phi} = (\tau, s_1, s_2, \ldots, s_n)$. Furthermore, the mean function is supposed to be linearly related to $\boldsymbol{\beta}$, that is,

$$\boldsymbol{m}(\mathbf{X}) = \mathbf{F}(\mathbf{X})\boldsymbol{\beta}, \tag{5}$$

where $\mathbf{F}(\mathbf{X}) \in \mathbb{R}^{m \times p}$ denotes the regressor matrix, whose i^{th} row, $\boldsymbol{f}_i(\mathbf{x})^T$, is called the basis function. Some common basis functions proposed in the literature are the linear basis function given by $\boldsymbol{f}(\mathbf{x}) = [1, x_1, \ldots, x_n]^T$ and the quadratic basis function given by $\boldsymbol{f}(\mathbf{x}) = [1, x_1, \ldots, x_n, x_1^2, \ldots, x_n^2]^T$.

To estimate $\boldsymbol{m}(\mathbf{X})$, we first estimate $\boldsymbol{\beta}$ and then use (5). To this end, we solve for the maximum likelihood estimator of the hyperparameters of the covariance function, $\boldsymbol{\phi}$, and of the mean function, $\boldsymbol{\beta}$, and the hyperparameter variance, σ_n^2, given the observation vector, $\mathbf{y}$, by maximizing the log-likelihood

function expressed as

$$\log(\mathcal{L}(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\phi}, \sigma_n^2)) = -\frac{1}{2}\boldsymbol{r}^T\mathbf{R}^{-1}\boldsymbol{r} - \frac{m}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{R}|\,, \tag{6}$$

where $\boldsymbol{r} = \mathbf{y} - \boldsymbol{m}(\mathbf{X})$ is the residual vector and $\mathbf{R} = \boldsymbol{k}(\mathbf{X}, \mathbf{X} \mid \boldsymbol{\phi}) + \sigma_n^2\mathbf{I}_m$ is the covariance matrix of $\mathbf{y}$. The maximum likelihood estimation problem can be cast as

$$(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\phi}}, \widehat{\sigma}_n^2) = \underset{(\boldsymbol{\beta}, \boldsymbol{\phi}, \sigma_n^2) \in \mathbb{R}^{p+1}\mathbb{R}^{n+1}\mathbb{R}^{+*}}{\operatorname{argmin}} - \log \mathcal{L}(\mathbf{y} \mid \boldsymbol{\beta}, \boldsymbol{\phi}, \sigma_n^2). \tag{7}$$

To solve (7), we first estimate $\boldsymbol{\beta}$ in the generalized least-squares sense, which is expressed as

$$\widehat{\boldsymbol{\beta}} = (\mathbf{F^T R^{-1} F})^{-1}\mathbf{F^T R^{-1} y}. \tag{8}$$

Then, we estimate the hyperparameters associated with the kernel function $\boldsymbol{\phi}$ and variance σ_n^2 by maximizing the $\boldsymbol{\beta}$-profile likelihood over $(\boldsymbol{\phi}, \sigma_n^2)$ as follows:

$$(\widehat{\boldsymbol{\phi}}, \widehat{\sigma}_n^2) = \underset{\boldsymbol{\phi}}{\operatorname{argmin}} - \log \mathcal{L}(\mathbf{y} \mid \widehat{\boldsymbol{\beta}}, \boldsymbol{\phi}, \sigma_n^2). \tag{9}$$

The gradient-based algorithm used to solve (9) is provided in Section 3.

Up to now, we have estimated all the hyperparameters of the GP model. We next validate this model. Let $(\mathbf{y}, \mathbf{X})$ be a collection of raw metered values of the system, called the training data set. Let $(\mathbf{X}^*, \mathbf{y}^*)$ be a new raw metered data set used to evaluate the performance of the GP model, called the validation data set. The training and validation data sets are assumed to follow a joint multivariate Gaussian distribution given by

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{y}^* \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{m}(\mathbf{X}) \\ \boldsymbol{m}(\mathbf{X}^*) \end{bmatrix}, \begin{bmatrix} \mathbf{R} & \mathbf{C}(\mathbf{X}^*) \\ \mathbf{C}(\mathbf{X}^*)^T & \mathbf{V}(\mathbf{X}^*) \end{bmatrix} \right), \tag{10}$$

where $\mathbf{C}(\mathbf{X}^*) = \boldsymbol{k}(\mathbf{X}, \mathbf{X}^*)$ and $\mathbf{V}(\mathbf{X}^*) = \boldsymbol{k}(\mathbf{X}^*, \mathbf{X}^*)$. Using the Bayes theorem, we infer the posterior distribution of the quantities that are of interest to us, which are $\mathcal{H}(\mathbf{X}^*)$, conditioned upon the training data set, $(\mathbf{y}, \mathbf{X})$, that is,

$$\mathcal{H}(\mathbf{X}^*) \mid \mathbf{X}^*, \mathbf{y}, \mathbf{X} \sim \mathcal{N}\left(\boldsymbol{\mu}^*(\mathbf{X}^*), \boldsymbol{\Sigma}^*(\mathbf{X}^*)\right). \tag{11}$$

The maximum a posteriori (MAP) estimators of $\boldsymbol{\mu}^*$ and $\boldsymbol{\Sigma}^*$, are respectively given by

$$\widehat{\boldsymbol{\mu}}^*(\mathbf{X}^*) = \widehat{\boldsymbol{m}}(\mathbf{X}^*) + \widehat{\mathbf{C}}(\mathbf{X}^*)^T\widehat{\mathbf{R}}^{-1}\boldsymbol{r}; \tag{12}$$

$$\widehat{\boldsymbol{\Sigma}}^*(\mathbf{X}^*) = \widehat{\mathbf{V}}(\mathbf{X}^*) - \widehat{\mathbf{C}}(\mathbf{X}^*)^T\widehat{\mathbf{R}}^{-1}\widehat{\mathbf{C}}(\mathbf{X}^*), \tag{13}$$

where $\widehat{\mathbf{R}} = \boldsymbol{k}(\mathbf{X}, \mathbf{X} \mid \widehat{\boldsymbol{\phi}}) + \widehat{\sigma}_n^2\mathbf{I}_m$, $\widehat{\mathbf{C}}(\mathbf{X}^*) = \boldsymbol{k}(\mathbf{X}, \mathbf{X}^* \mid \widehat{\boldsymbol{\phi}})$, $\widehat{\mathbf{V}}(\mathbf{X}^*) = \boldsymbol{k}(\mathbf{X}^*, \mathbf{X}^* \mid \widehat{\boldsymbol{\phi}})$. The shortcomings of the conventional hyperparameter estimators are: (i)

the non-robustness to outliers of the mean function hyperparameter estimator, $\widehat{\boldsymbol{\beta}}$ and (ii) the deviations of the assumed prior probabilities in the likelihood function from the actual ones.

2.2 The SHGM Estimator

Huber (1964) initiated a new class of maximum-likelihood-type estimators, called M-estimators, which minimize objective functions of the form

$$\mathcal{J}(\boldsymbol{\beta}) = \sum_{i=1}^{m} \rho\left(\frac{r_i}{s^*}\right), \tag{14}$$

where $\rho\left(r_i/s^*\right)$ is a function of the standardized residuals. The Huber M-estimator is defined by a convex ρ-function given by

$$\rho(r_i) = \begin{cases} \frac{r_i^2}{2}, & \text{for } |r_i| \leq b; \\ b \mid r_i \mid -\frac{b^2}{2}, & \text{for } |r_i| > b, \end{cases} \tag{15}$$

where b is the threshold, typically set to 1.5 to get a high statistical efficiency at the Gaussian distribution. The Shweppe-type GM estimator (SHGM) minimizes a weighted Huber objective function defined as

$$J(\boldsymbol{\beta}) = \sum_{i=1}^{m} w_i^2 \rho\left(\frac{r_i}{s^* w_i}\right), \tag{16}$$

where $s^* = 1.4826\ b_m \text{median} \mid \boldsymbol{r} \mid$ is the robust estimator of the scale of the residuals when the error covariance is unknown (Mili et al. (1992)) and where $b_m = 1 + 5/(m-n)$ accounts for the dimensionality correction factor. To bound the influence of the outliers, the weights $w(\boldsymbol{f}_i)$ in (16) are defined in such a way that they will decrease from 1 to 0 as the degrees of outlyingness of the corresponding data point get larger. This is performed by defining the weights as functions of the projection statistics as described in Section 3. As for the SHGM estimator, it is a solution of

$$\sum_{i=1}^{m} w_i \boldsymbol{f}_i \psi(r_{S_i}) = 0. \tag{17}$$

where $\psi(r_{S_i}) = \frac{\partial \rho(r_{S_i})}{\partial r_{S_i}}$ is known as the ψ-function. The Huber ψ-function is expressed as

$$\psi(r_{S_i}) = \begin{cases} r_{S_i}, & \text{for} |r_{S_i}| \leq b; \\ b\ sign(r_{S_i}), & \text{for} |r_{S_i}| > b, \end{cases} \tag{18}$$

and where the standardized residuals are given by $r_{S_i} = r_i/(s^* w_i)$. Since $\psi(r_{S_i})$ is a non-linear function, we resort to an iteratively reweighted least-squares (IRLS) algorithm (Beaton et al. (1974)). Using (17), we get

$$\sum_{i=1}^{m} q\left(\frac{r_i}{w_i s}\right) \boldsymbol{f}_i r_i = 0, \tag{19}$$

where $q(r_{S_i}) = \frac{\mathrm{PS}i(r_{S_i})}{r_{S_i}}$ is known as the weight function given by

$$q(r_{S_i}) = \begin{cases} 1, & r_i \leq b \\ \frac{b \ \mathrm{sign}(r_{S_i})}{r_{S_i}}, & \text{otherwise} \end{cases}. \tag{20}$$

Finally, substituting $r_i = y_i - \boldsymbol{f}_i^T \boldsymbol{\beta}$ in (19) and putting the results in a matrix form, we get

$$\mathbf{F}^T \mathbf{Q}(\mathbf{y} - \mathbf{F}\boldsymbol{\beta}) = \mathbf{0}, \tag{21}$$

where $\mathbf{Q}$ is called the weight matrix given by $\mathbf{Q} = \mathrm{diag}(\boldsymbol{q}(\boldsymbol{r}_S))$. Since $\mathbf{Q}$ is a function of $\boldsymbol{\beta}$, we solve (21) for $\boldsymbol{\beta}$ using the IRLS algorithm expressed as

$$\boldsymbol{\beta}^{(k+1)} = (\mathbf{F}^T \mathbf{Q}^{(k)} \mathbf{K}^{-1} \mathbf{F})^{-1} \mathbf{F}^T \mathbf{Q}^{(k)} \mathbf{K}^{-1} \mathbf{y}, \tag{22}$$

where $\mathbf{K}$ denotes the covariance matrix of errors. This algorithm is preferred to Newton's algorithm due to its simplicity and its numerical stability.

3 Robust Observational Data-Driven Process Model

In this section, we introduce the projection statics on which the weights for the SHGM estimator are based. We then develop the RODP model and analyze its robustness with the help of a numerical model.

The projection statistics are defined as the maxima of the standardized projection distances obtained by projecting the point cloud on the directions that originate from the coordinate-wise median and that pass through each of the data points $\boldsymbol{f}_i$ (see, Mili (1992)). Formally, we have

$$\mathrm{PS}_i = \underset{||\boldsymbol{u}||=1}{max} \frac{\boldsymbol{f}_i^T \boldsymbol{u} - \underset{j}{\mathrm{median}}(\boldsymbol{f}_j^T \boldsymbol{u})}{1.4826 \ \underset{k}{\mathrm{median}} \ | \ \boldsymbol{f}_k^T \boldsymbol{u} - \underset{j}{\mathrm{median}}(\boldsymbol{f}_j^T \boldsymbol{u}) \ |}, \tag{23}$$

where PS_i is the projection statistic associated with $\boldsymbol{f}_i$. The projection statistics are fast to calculate; but this comes at the expense of a loss of affine equivariance, which is not a shortcoming for our method because they are used

as a robust diagnostic tool to identify outliers. We define the weights for the SHGM estimator in (16) based on projection statistics as

$$w(\boldsymbol{f}_i) = \begin{cases} 1, & \text{for } |r_{Si}| \leq c, \\ \frac{c}{\text{PS}_i^2}, & \text{for } |r_{Si}| > c, \end{cases} \tag{24}$$

which downweight both the bad leverage points and the vertical outliers but not the good leverage points. The threshold parameter, c, is set to the 97.5 percentile of the chi-squared distribution with n degrees of freedom, that is, $\chi^2(0.975, n)$ (see, Zomeren et al. (1991)).

3.1 Development of the RODP

The derivation of the RODP model is based on robust statistics initiated by Huber (1964) and the robustness concepts proposed by Hampel et al. (1986). First, the mean function of the GP model given by (5) is re-written as

$$\boldsymbol{\zeta}(\mathbf{X}) = \mathbf{F}(\mathbf{X})\boldsymbol{\beta} + \boldsymbol{e}(\mathbf{X}). \tag{25}$$

Then, the projection statistics are calculated on the regressor matrix, $\mathbf{F}(\mathbf{X})$, and the outliers are flagged as the data points whose corresponding standardized residual magnitudes, $\mid r_{Si} \mid$, are greater than c. After identifying the outliers, we downweight them via the weights given by (24). In our RODP model framework, we aim to estimate the mean function hyperparameter, $\boldsymbol{\beta}$, in a robust manner using the SHGM estimator. The estimate, $\hat{\boldsymbol{\beta}}$, is the solution of the IRLS algorithm given by (22), which is expressed as

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{\phi}, \sigma_n^2) = \left[\mathbf{F}(\mathbf{X})^T \mathbf{Q} \left[\boldsymbol{k}(\mathbf{X}, \mathbf{X} \mid \boldsymbol{\phi}) + \sigma_n^2 \mathbf{I}_m\right]^{-1} \mathbf{F}\right]^{-1} \mathbf{F}(\mathbf{X})^T \mathbf{Q} \left[\boldsymbol{k}(\mathbf{X}, \mathbf{X} \mid \boldsymbol{\phi}) + \sigma_n^2 \mathbf{I}_m\right]^{-1} \mathbf{y}. \tag{26}$$

Using the observation set, $(\mathbf{y}, \mathbf{X})$, we now estimate the hyperparameters, $\boldsymbol{\eta} = (\widehat{\boldsymbol{\beta}}, \boldsymbol{\phi}, \sigma_n^2)$, of the RODP model as

$$\widehat{\boldsymbol{\eta}} = \underset{(\widehat{\boldsymbol{\beta}}, \boldsymbol{\phi}, \sigma_n^2) \in \mathbb{R}^{p+1} \mathbb{R}^n \mathbb{R}^{+*}}{\arg\max} \log \mathcal{L}\left(\mathbf{y} \mid \mathbf{X}, \widehat{\boldsymbol{\beta}}, \boldsymbol{\phi}, \sigma_n^2\right). \tag{27}$$

Proceeding with the notations specified in Section 2.1, the log-likelihood function $\mathcal{L}$ is given by

$$\log \mathcal{L}\left(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\phi}, \sigma_n^2\right) = -\frac{1}{2}(\mathbf{y} - \mathbf{F}\boldsymbol{\beta})^T \mathbf{R}^{-1} (\mathbf{y} - \mathbf{F}\boldsymbol{\beta}) - \frac{m}{2} \log(2\pi) - \frac{1}{2} \log |\mathbf{R}|. \tag{28}$$

Substituting $\widehat{\boldsymbol{\beta}}$ into (28), we obtain the estimates of $(\boldsymbol{\phi}, \sigma_n^2)$ as

$$\left(\widehat{\boldsymbol{\phi}}, \widehat{\sigma}_n^2\right) = -\underset{(\boldsymbol{\phi}, \sigma_n^2)}{\arg\min}\, \Gamma(\boldsymbol{\phi}, \sigma_n^2), \tag{29}$$

where the Γ-function is defined as

$$\Gamma(\boldsymbol{\phi}, \sigma_n^2) = \log |\mathbf{R}|. \tag{30}$$

We solve (30) by applying a gradient-based optimizer (see, Rassmussen and Willaims (2006)) as follows. We begin by estimating $\boldsymbol{\phi}$ using the gradient-descent algorithm given by

$$\boldsymbol{\phi}^{(k+1)} = \boldsymbol{\phi}^{(k)} - \gamma \nabla_{\boldsymbol{\phi}} \Gamma(\boldsymbol{\phi}, \sigma_n^2),$$

where γ is a step length. The gradients of Γ-function with respect to $\boldsymbol{\phi}$, $\nabla_{\boldsymbol{\phi}} \Gamma(\boldsymbol{\phi}, \sigma_n^2)$, are given by

$$\frac{\partial \Gamma(\boldsymbol{\phi}, \sigma_n^2)}{\partial \boldsymbol{\phi}} = \left[\frac{\partial \Gamma(\boldsymbol{\phi}, \sigma_n^2)}{\partial \tau}, \frac{\partial \Gamma(\boldsymbol{\phi}, \sigma_n^2)}{\partial s_1}, \ldots, \frac{\partial \Gamma(\boldsymbol{\phi}, \sigma_n^2)}{\partial s_n} \right],$$

$$\frac{\partial \Gamma(\boldsymbol{\phi}, \sigma_n^2)}{\partial \tau} = (2\tau) \mathrm{tr} \left((\mathbf{R})^{-1} \mathbf{A}(\mathbf{X}, \mathbf{X}) \right),$$

$$\frac{\partial \Gamma(\boldsymbol{\phi}, \sigma_n^2)}{\partial s_l} = (-2) \mathrm{tr} \left((\mathbf{R})^{-1} \boldsymbol{S}(l) \right),$$

where

$$\mathbf{A}(\mathbf{X}, \mathbf{X}) = \begin{bmatrix} R_c(\mathbf{x}_1, \mathbf{x}_1) & \ldots & R_c(\mathbf{x}_1, \mathbf{x}_m) \\ \vdots & \ddots & \vdots \\ R_c(\mathbf{x}_m, \mathbf{x}_1) & \ldots & R_c(\mathbf{x}_m, \mathbf{x}_m) \end{bmatrix},$$

and

$$\boldsymbol{S}(l) = \begin{bmatrix} \frac{(\mathrm{x}_{1l} - \mathrm{x}_{1l})^2}{s_l^3} R_c(\mathbf{x}_1, \mathbf{x}_1) & \ldots & \frac{(\mathrm{x}_{1l} - \mathrm{x}_{ml})^2}{s_l^3} R_c(\mathbf{x}_1, \mathbf{x}_m) \\ \vdots & \ddots & \vdots \\ \frac{(\mathrm{x}_{ml} - \mathrm{x}_{1l})^2}{s_l^3} R_c(\mathbf{x}_m, \mathbf{x}_1) & \ldots & \frac{(\mathrm{x}_{ml} - \mathrm{x}_{ml})^2}{s_l^3} R_c(\mathbf{x}_m, \mathbf{x}_m) \end{bmatrix},$$

for $l = 1, 2, \ldots, n$.

Then, we estimate σ_n^2 by substituting $\widehat{\boldsymbol{\phi}}$ in the Γ-function given by (30), which yields

$$\widehat{\sigma}_n^2 = -\arg\min_{(\sigma_n^2)} \Gamma(\sigma_n^2, \widehat{\boldsymbol{\phi}}).$$

We solve for $\widehat{\sigma}_n^2$ using the gradient-descent algorithm given by

$$\sigma_n^{2\,(k+1)} = \sigma_n^{2\,(k)} - \alpha \nabla_{\sigma_n^2} \Gamma(\boldsymbol{\phi}, \sigma_n^2),$$

where α is a step length. The gradient of the Γ-function with respect to σ_n^2, $\nabla_{\sigma_n^2} \Gamma(\boldsymbol{\phi}, \sigma_n^2)$, is given by

$$\frac{\partial \Gamma(\boldsymbol{\phi}, \sigma_n^2)}{\partial \sigma_n^2} = tr(\mathbf{R})^{-1}.$$

With the estimates $(\widehat{\boldsymbol{\phi}}, \widehat{\sigma}_n^2)$, we now update $\widehat{\boldsymbol{\beta}}_{\mathrm{RODP}}(\boldsymbol{\phi}, \sigma_n^2)$ to get $\widehat{\widehat{\boldsymbol{\beta}}}_{\mathrm{RODP}}$. We obtain the final estimate of $\widehat{\boldsymbol{\eta}}$ by solving (26) and (29) in an iterative manner. The algorithm for calculating $\widehat{\boldsymbol{\eta}}$ is outlined in Table 1.

The mean function of the RODP model that acts as a robust surrogate is given by

$$\widehat{\boldsymbol{\mu}}^*(\mathbf{X}^*) = \mathbf{F}(\mathbf{X}^*) \widehat{\widehat{\boldsymbol{\beta}}} + \widehat{\mathbf{C}}(\mathbf{X}^*)^T \widehat{\mathbf{R}}^{-1} \boldsymbol{r}. \tag{31}$$

The covariance matrix of the RODP model is given by (13) using the newly robust hyperparameters of the covariance function, $(\widehat{\boldsymbol{\phi}}, \widehat{\sigma_n^2})$. We now make the desired predictions using the computationally efficient robust model given by (31).

Algorithm 1 Estimation of the RODP Hyperparameters $\{\boldsymbol{\beta}, \boldsymbol{\phi}, \sigma^2\}$

Require: m training samples $(\mathbf{y}, \mathbf{X})$, projection statistics **PS**.

if $m > 5n$ **then**
  $w(\boldsymbol{f}_i) = \min(1, \frac{c}{\mathrm{PS}_i^2})$
else
  $w(\boldsymbol{f}_i) = \min(1, \frac{c}{\mathrm{PS}_i})$
end if
Initialize $\boldsymbol{\beta}_0 = \boldsymbol{\beta}_{wls}, \boldsymbol{\phi}_0 = [1, \ldots, 1]^T, \sigma_0^2 = 0.001$
while $error > 0.001$ **do**
  if $r_{S_i} \leq b$ **then**
    $q(r_{S_i}) = 1$
  else$r_{S_i} > b$
    $q(r_{S_i}) = \frac{b\ sign(r_{S_i})}{r_{S_i}}$
  end if
  $\mathbf{Q} = diag(\boldsymbol{q}(\boldsymbol{r}_S))$
  $\boldsymbol{\beta}^{k+1} = (\mathbf{F}^T\mathbf{Q}^{(k)}\mathbf{R}^{-1}\mathbf{F})^{-1}\mathbf{F}^T\mathbf{Q}^{(k)}\mathbf{R}^{-1}\mathbf{y}$
end while
$\widehat{\boldsymbol{\beta}} = (\mathbf{F}^T\mathbf{Q}^{(k)}\mathbf{R}^{-1}\mathbf{F})^{-1}\mathbf{F}^T\mathbf{Q}^{(k)}\mathbf{R}^{-1}\mathbf{y}$
$\left(\widehat{\boldsymbol{\phi}}, \widehat{\sigma}_n^2\right) = -\underset{(\boldsymbol{\phi}, \sigma_n^2)}{\arg\min}\, \Gamma(\boldsymbol{\phi}, \sigma_n^2)$
Iterate Steps 16 and 17 until convergence, e.g. $||\boldsymbol{r}|| \leq 0.001$, to obtain the final hyperparameter estimates, $\widehat{\widehat{\boldsymbol{\beta}}} = \hat{\boldsymbol{\beta}}(\widehat{\boldsymbol{\phi}}, \widehat{\sigma}_n^2), \widehat{\boldsymbol{\phi}}$, and $\widehat{\sigma}_n^2$

3.2 Robustness and Efficiency of the RODP model Model

We assess the robustness of the RODP model using the three robustness concepts proposed by Hampel (1986), namely the influence function, the breakdown point, and the statistical efficiency. The reader is referred to Appendix A for further details about the robustness concepts. Let us consider an ε-contaminated model $G(r) = (1-\varepsilon)\Phi(r) + \varepsilon\Theta(r)$ where $\Phi(r)$ is the cumulative Gaussian distribution function and $\Theta(r)$ is the unknown distribution function of the residuals.

Theorem: The influence function IF of the SHGM estimator at the probability distribution, G, is defined as

$$\mathrm{IF}(r_{S_i}, \boldsymbol{f}_i;\ \Phi) = \frac{\psi(r_S)}{E_\Phi[\psi'(r_S)]}(\mathbf{F}^T\mathbf{F})^{-1}\boldsymbol{f}_i w(\boldsymbol{f}_i), \tag{32}$$

where r_{S_i}, for $i = 1, 2, \ldots, m$, are the standardized residuals.
The reader is referred to Appendix C for the proof. As for the asymptotic covariance matrix of the $\boldsymbol{\beta}$ estimator (see, Gandhi et al. (2010)), it is expressed as

$$\widehat{\boldsymbol{\varsigma}} = E_\Phi[\mathrm{IF}\ \mathrm{IF}^T], \tag{33}$$

which yields,

$$\widehat{\boldsymbol{\varsigma}} = \frac{E_\Phi[\psi^2(r_S)]}{\left(E_\Phi[\psi'(r_S)]\right)^2}\left(\mathbf{F}(\mathbf{X})^T\mathbf{F}(\mathbf{X})\right)^{-1}\left(\mathbf{F}(\mathbf{X})^T\mathbf{Q}\mathbf{F}(\mathbf{X})\right)\left(\mathbf{F}(\mathbf{X})^T\mathbf{F}(\mathbf{X})\right)^{-1}. \tag{34}$$

Yohai et al. (1995) showed that the projection statistics attain the maximum breakdown point given by $[(m-n-1)/2]/m$. The trade-off between the breakdown

point and the statistical efficiency at the Gaussian distribution of the RODP model is addressed by taking advantage of the SHGM estimator. The SHGM estimator reduces to an ℓ_2-norm estimator for small standardized residuals and to the ℓ_1-norm estimator for larger ones. This estimator has a bounded influence function and a positive breakdown point, which indicates that it is robust to outliers. Furthermore, it exhibits good statistical efficiency for a broad range of thick-tail probability distributions.

4 CASE STUDIES

In this section, we first demonstrate the robustness and applicability of the proposed model to thick-tail distributions on an artificial two-input model. We then validate the proposed RODP model using the diagnostics proposed by Oakley and O'Hagan (2004) and compare the results to traditional GP model predictions. We finally apply the proposed model to two real-world problems, namely, (i) to carry out a time-series prediction of the wind speed by processing Chichester Harbour, UK location weather data; (ii) to perform stochastic power flow predictions by processing real-world power injection measurements of the IEEE 39-bus power system that is heavily integrated with renewable energy resources.

4.1 Case Study 1: Artificial Model

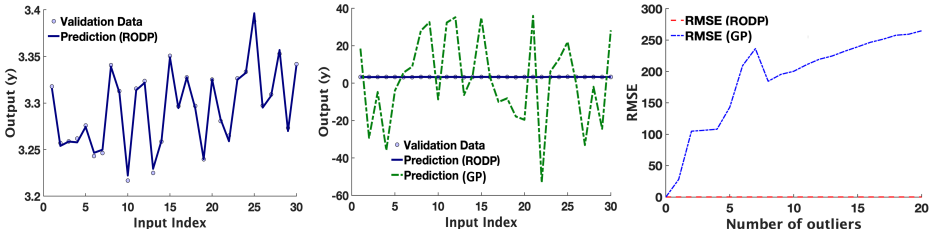
The following artificial model is considered with a linear basis function $\boldsymbol{f}_i = [1,\ x_{1i},\ x_{2i}]^T, \quad i = \{1, \ldots, N\}$:

$$\mathcal{H}(x_1, x_2) = \left(1 - \exp\left(\frac{-1}{x_1}\right)\right)\left(\frac{57x_1^3 + 19x_1^2 + 20x_1 + 60}{10x_2^3 + 48x_2^2 + 4x_2 + 18}\right). \tag{35}$$

A sample size of $N = 110$ points constitutes the data set; out of them, $m = 80$ data points are used as a training data set, contained in $(\mathbf{X}, \mathbf{y})$, to offer enough redundancy for predictions at $m_t^* = 30$ test data set, contained in $(\mathbf{X}^*, \mathbf{y}^*)$, called the validation data set.

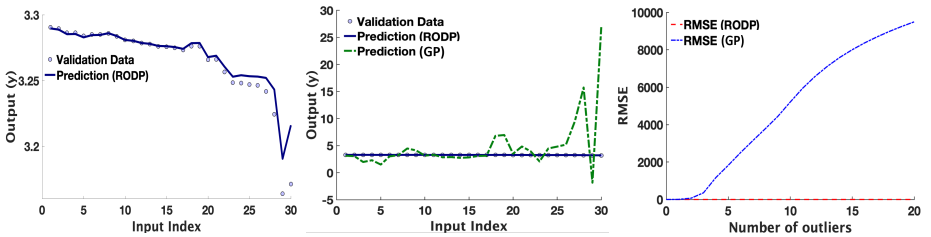
4.1.1 Applicability to a Wide Range of Thick Tail Probability Distributions

We exemplify the applicability of the proposed model to the Gaussian (Figure 1), the Laplace (Figure 2), and the Student-t (Figure 3) distributions by comparing the prediction values between the GP and the RODP model. The degrees of freedom considered for the Student-t input variables are two. For each case, we add up to 25% of outliers with large magnitudes to the training data, which consist of bad leverage points (outliers both in $\mathbf{y}$ and $x_i, i = 1, 2$)) and vertical outliers (outliers only in $\mathbf{y}$, but not in $\mathbf{x}$). We also add good leverage points, i.e. $\mathbf{y}$ values obtained via artificial model runs at the input variables with outliers only in $\mathbf{x}_i, i = 1, 2$. The RODP model results are accomplished by setting the parameter b of the Huber ρ-function to 1.5 for the Gaussian and the Student-t distribution and to 0.01 for the Laplace distribution. We notice significant deviated prediction values from the true values using the GP model with the addition of a single bad leverage point while the RODP model exhibits a low root mean square error (RMSE) for all the three probability distributions (see Figures 1(b), 2(b), and 3(b)). The RMSE comparisons



(a) (b) (c)

Fig. 1: Results for the case of the Gaussian distribution. The inputs and error are sampled as $X_1 \sim \mathcal{N}(50, 10), X_2 \sim \mathcal{N}(20, 4), \epsilon \sim \mathcal{N}(0, 0.001)$; (a) prediction values obtained from the RODP model with 25% of outliers in the training data; (b) comparison of the prediction values provided by the RODP model and the GP model trained on data with a single outlier; and (c) comparison of the RMSE obtained using the RODP and the GP model.



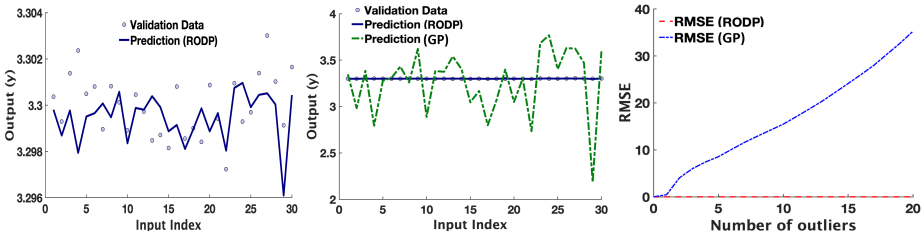
(a) (b) (c)

Fig. 2: Results for the case of the Laplace distribution. The inputs and error are sampled as $X_1 \sim \mathcal{L}(15, 2), X_2 \sim \mathcal{L}(10, 2), \epsilon \sim \mathcal{L}(0, 0.001)$, respectively; (a) RODP model prediction values; (b) comparison of the predictions obtained using the RODP and the GP model with a single outlier in the training data; and (c) the RMSE plot of the RODP and the GP model with up to 25% of outliers in the training data set.

reflect similar conclusions when the number of outliers are increased (see Figures 1(c), 2(c), and 3(c)).

4.1.2 Graphical Diagnostics Applied to the RODP and the GP model

The graphical diagnostic techniques considered here are based on the standardized residuals $\mathbf{D}_i^I(\mathbf{y}^*) = \frac{y_i^* - \widehat{\mu}_i^*(\mathbf{x}_i^*)}{\sqrt{V(\mathbf{x}_i^*)}}$ and the Cholesky errors $\mathbf{D}^G(\mathbf{y}^*) = \mathbf{G}^{-1}\left(\mathbf{y}^* - \widehat{\boldsymbol{\mu}}^*(\mathbf{X}^*)\right)$ developed by Oakley and O'Hagan (2004), where $\mathbf{G}$ is the lower triangular matrix using the Cholesky decomposition of $\mathbf{V}(\mathbf{X}^*)$. They are plotted in Figures 4 to 6 for the case of 25% outliers in the input data sets. Here, the training data and validation data points are obtained using Latin hypercube sampling to fill



(a) (b) (c)

Fig. 3: Results for the case of the Student-t distribution. The inputs are sampled as $X_i \sim$ Student-t($\nu = 2$) for $i = 1, 2$; (a) predicted values using the RODP model with 25% of outliers in the training data set; (b) comparison of the predicted values using the RODP and the GP model for a single outlier in the training data set; and (c) the RMSE plot for the RODP and the GP model for an increasing percentage of outliers up to 25% in the training data set.

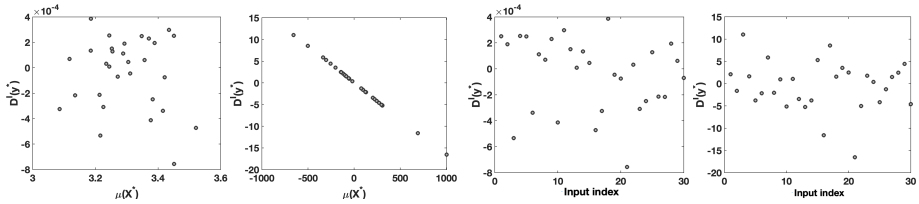
the design space, $x_i \in (0, 1]$ for $i = 1, 2$, except for 20 training data points associated with x_1 that accounts for bad leverage points with values outside the range $(0, 1]$. The hyperparameters of the mean function, $\boldsymbol{\beta}$, estimated using the RODP and the GP model are valued as $\widehat{\boldsymbol{\beta}}_{\text{RODP}} = [-5.1427,\ 1.3059,\ -2.1056]$ and $\widehat{\boldsymbol{\beta}}_{GP} = [1.8390 \times 10^5,\ 3.2079 \times 10^5,\ -1.04882 \times 10^4]$, respectively. The estimated kernel function hyperparameters, ($\boldsymbol{\phi}$), of the RODP model are $\widehat{\boldsymbol{\phi}}_{\text{RODP}} = [2.9164 \times 10^4,\ 15.2861,\ 4.8654]$ and that of the GP model are $\widehat{\boldsymbol{\phi}}_{\text{GP}} = [2.9164 \times 10^4,\ 4.9459,\ 3.0814]$.

We observe from Figures 4(a) and (b) that the standardized errors of the RODP model are smaller than those of the GP model. Interestingly, these results are at odd with the diagnostics based on the standardized errors and the Cholesky errors plotted versus the index (see Figures 4(d) and 5(b), respectively) and the standardized errors versus the input 1 (see Figure 6(b)) for the GP model display a constant variation band, which suggest a well estimated mean function. However, the plot of the standardized errors versus the mean prediction (see Figure 4(b)) indicates an unbounded linear trend. Thus, the Mahalanobis distance-based diagnostics are unreliable in presence of outliers. This raises the need for robust diagnostics to validate the model.

4.2 Real Applications

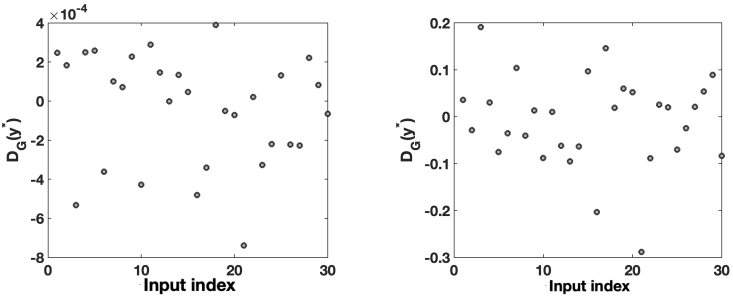
4.2.1 Case Study 2: Short Term Wind Speed and Water Depth Forecast

Substantial computational resources are required to perform weather forecasting, which is governed by thermodynamics and fluid dynamics equations. Classical autoregressive time-series models capture solely the linear dependencies in the time series. To overcome this problem, researchers in this area are proposing to use the GP non-parametric models, which allow us to estimate the prediction uncertainty and intervals while accounting for model non-linearity. Zhang et al. (2016) proposed GP regression models for both the point and interval predictions for the wind speed forecast. They compared the conventional artificial neural network (ANN) and the



(a) (b) (c) (d)

Fig. 4: (a) Standardized errors $\mathbf{D}^I(\mathbf{y}^*)$ versus the mean prediction $\widehat{\boldsymbol{\mu}}(\mathbf{X}^*)$ using the RODP model; (b) $\mathbf{D}^I(\mathbf{y}^*)$ versus the mean prediction $\widehat{\boldsymbol{\mu}}(\mathbf{X}^*)$ using the GP model; (c) $\mathbf{D}^I(\mathbf{y}^*)$ versus the input index using the RODP model; and (d) $\mathbf{D}^I(\mathbf{y}^*)$ versus the input index using the GP model.



(a) (b)

Fig. 5: Cholesky errors $\mathbf{D}^G(\mathbf{y}^*)$ versus input index for (a) the RODP model model and (b) the GP model.

support vector machine (SVM) methods for one-step-ahead forecasting and demonstrated that the GP predictions are more accurate. Among many methods proposed in the literature for short-term wind prediction, very few of them address the adverse effects of outliers. However, the measurement process of weather phenomena possesses unusual values that result from sensor malfunction, hardware error, power supply error, ambient environment change, and, in some rare circumstances, abnormal weather phenomena. It becomes necessary to replace an outlier detected in the measurement process with a value that is consistent with the bulk of the data set, a value that may not always be determined by an expert. In this paper, we propose to use projection statistics-based weights, which are fast to compute, to bring the outlier value closer to the rest of the observed data.

The second example we provide is based on real-time time-series data that are collected by a Cambermet sensor on the south coast of the UK. The sensors record up-to-date measurements of wind speed and water depth and other environmental variables, including wind direction, air temperature, sea temperature, and tide height (Roberts et al. (2013)). We leverage two data streams of measurements of wind speed and water depth with a resolution of 5 minutes to assess the performance of the RODP model. An auto-regressive model is employed to capture the overall

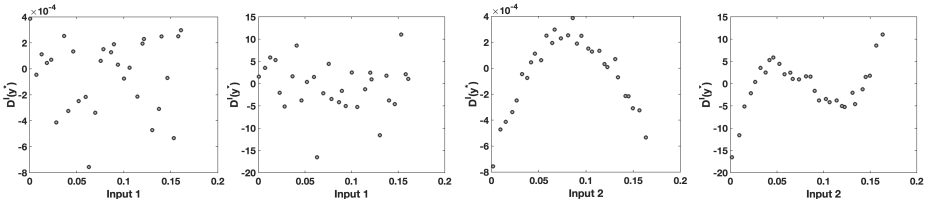


Fig. 6: Diagnostic technique based on $\mathbf{D}^I(\mathbf{y}^*)$ versus inputs: (a) $\mathbf{D}^I(\boldsymbol{y}^*)$ against Input 1 for the RODP; (b) $\mathbf{D}^I(\mathbf{y}^*)$ against Input 1 for the GP model; (c) $\mathbf{D}^I(\mathbf{y}^*)$ against Input 2 for the RODP model; and (d) $\mathbf{D}^I(\mathbf{y}^*)$ against Input 2 for the GP model.

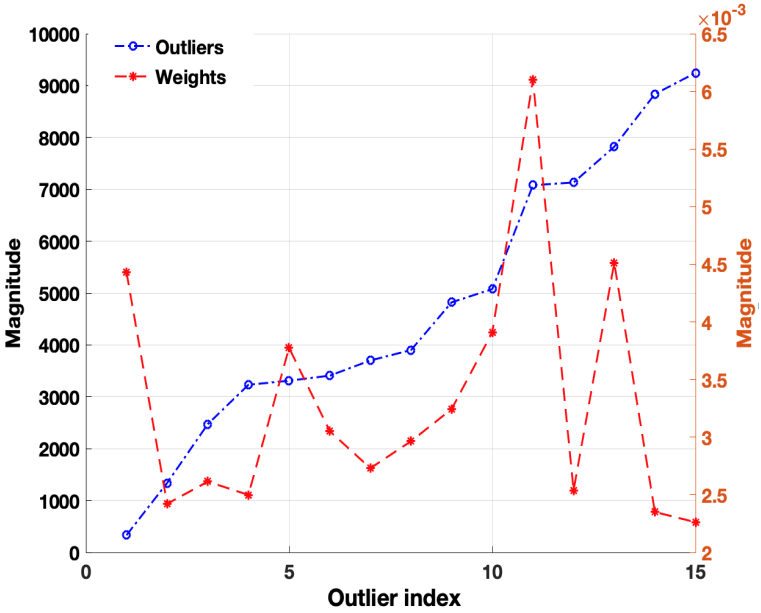
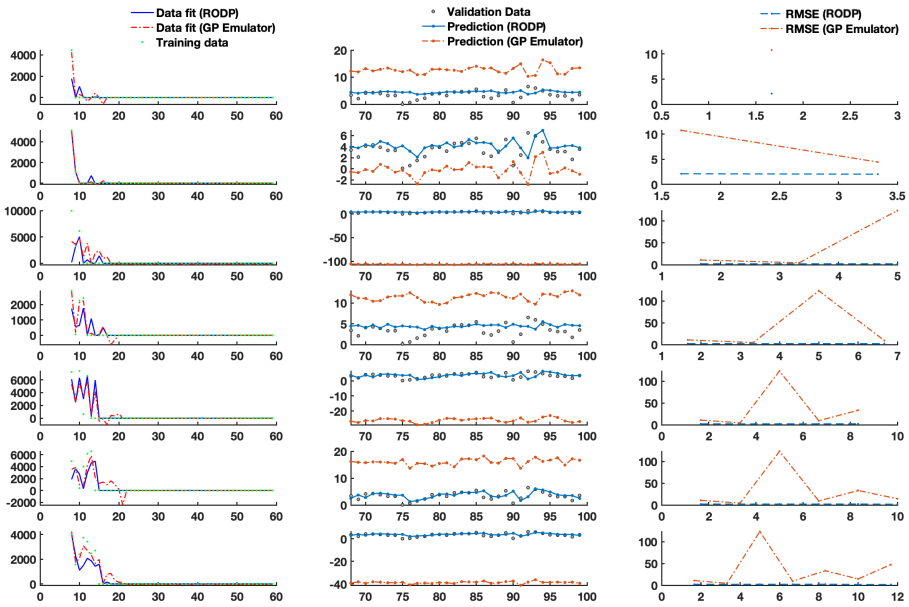
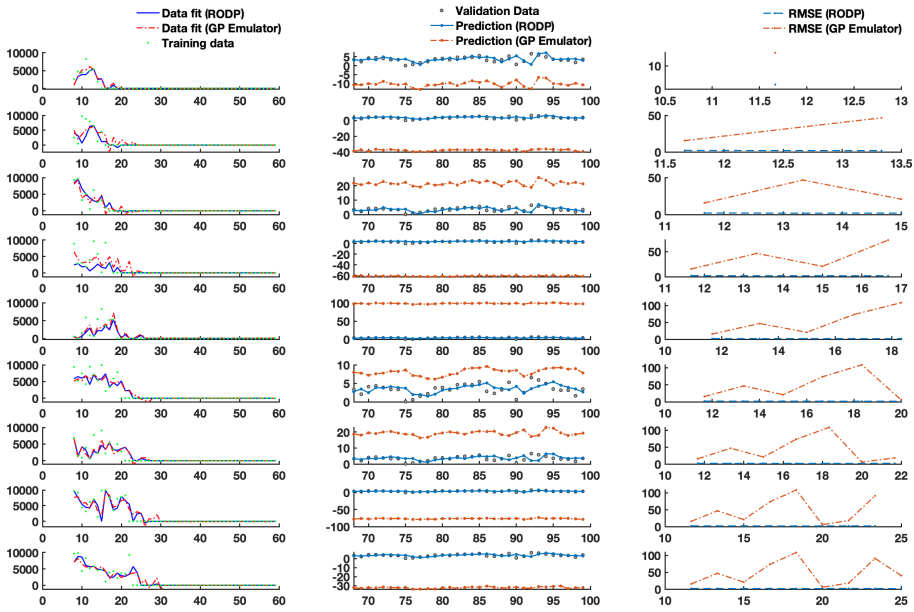


Fig. 7: Calculated weights and magnitudes of added outliers versus their index.

structure from the wind speed series. We evaluate the RODP model on three different cases of training data sets with 25% of outliers: (i) Case 1: the training set includes measurements recorded at $\boldsymbol{t} = [t_1, t_2, \ldots, t_{60}]^T$ that are used to forecast the wind speed over a short term of $m_t^* = 40$ min; (ii) Case 2: the training set includes measurements over $m = 188$ min for model validation over $m_t^* = 100$ min (see Figure 9); (iii) Case 3: the forecast of the water depth for the time period of $m_t^* = 100$ trained over $m = 188$ min. The magnitude of added outliers and associated weights are displayed in Figure 7. The comparison of the fitted data, prediction values, and RMSE obtained for Case 1 from the GP and the RODP model for an increasing percentage of outliers from 1% to 25% is displayed in Figure 8.



(a)



(b)

Fig. 8: Plots of the data fit (first column), prediction (second column), and RMSE (third column) of the wind speed forecast for Case 1 obtained using the RODP and the GP model when (a) 1% to 12% and (b) 13% to 25% of outliers are added to the training dataset.

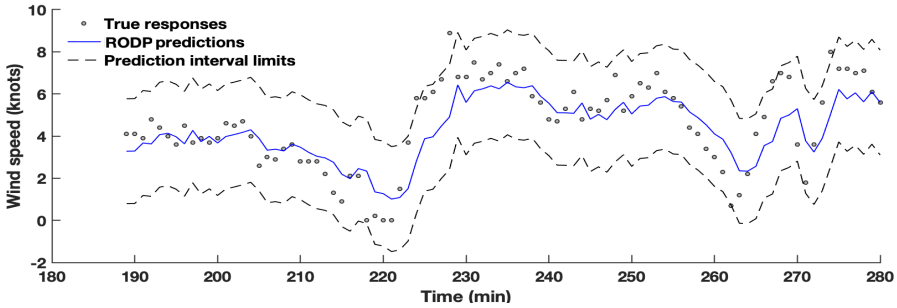


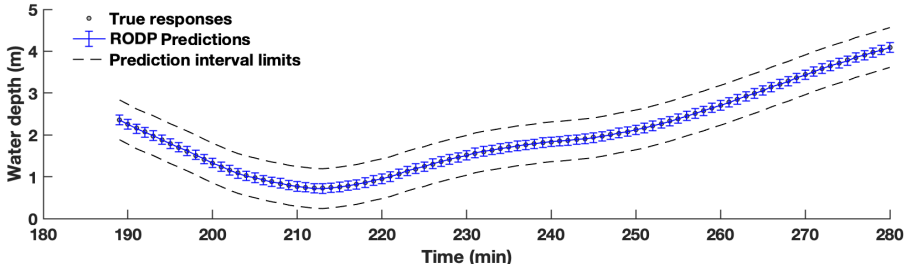
Fig. 9: Predictions for the next interval of 100 minutes trained on the first 188 minutes for Case 2.

The experimental results show that, as we increase the fraction of outliers in the training dataset, the GP model results start to deviate from the true values. In contrast, the RODP model consistently exhibits low RMSE. The results generate satisfactory prediction intervals for the water depth as displayed in Figure 10.

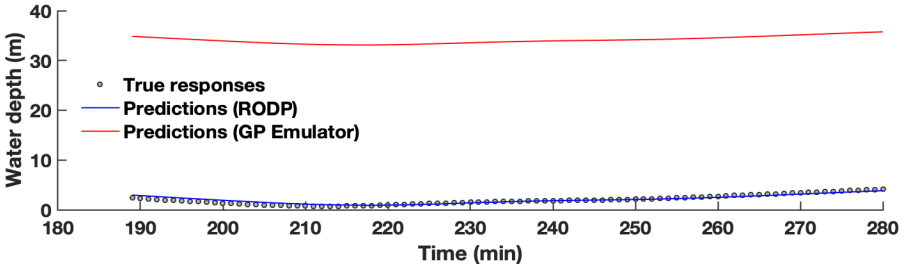
4.2.2 Case Study 3: Stochastic Power Flow

The nodal voltage magnitudes, phase angles, and line real and reactive power flow obtained from power flow calculations are important information for planning expansion and real-time economic operations of a power system. A number of power flow algorithms with good convergence properties have been proposed in the literature (see, Murty (2017)). As modern power systems are transitioning towards net-zero emissions, renewable energy sources (RES), electrified heat, and transport systems are inducing uncertainties in the power distribution system outputs due to their intermittences and stochastic variations. One such method to quantify the Uncertainty in modern power systems is using Monte Carlo simulations, which are computationally expensive. As measurement devices such as phasor measurement units are prevalent in the network, various data-driven methods have gained growing attention in the literature (see, Wang (2020), Wang et al. (2021)).

This paper considers a time-series stochastic power flow model for a radial power distribution system benchmarked as the IEEE 39-bus power distribution system. The model inputs are the active and reactive power injections at all buses $\mathbf{x}_i$ gathered in $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]^T$, while the model outputs are the nodal voltage magnitudes, $[y_{v1}, y_{v2}, \ldots, y_{vN}]^T$, and angles, $[y_{a1}, y_{a2}, \ldots, y_{aN}]^T$ at N instances. The power distribution system is integrated with four RESs, namely, one photovoltaic (PV) unit connected to Bus 38, whose output, real power injection, $\mathrm{P}_{t,G38}$, is measured at $\boldsymbol{t} = [t_1, t_2, \ldots, t_N]^T$ and three wind generation (WG) units connected to Buses $35, 36, 37$, whose output power injections, $\mathrm{P}_{t,G35}, \mathrm{P}_{t,G36}, \mathrm{P}_{t,G37}$, are measured. These measurements are taken 1 h apart for a total of $N = 424$ data points, of which the first $m = 400$ measurements constitute the training data. The RODP model allows us to design the model without the need to specify the probability distributions of the loads or of the PV and WG power outputs. For the stochastic power flow at test points $m_t^* = 24$, a total of 5000 samples are generated using Latin Hypercube Sampling from



(a)

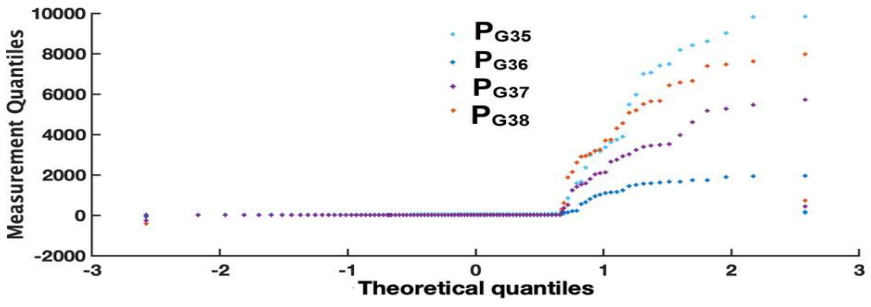


(b)

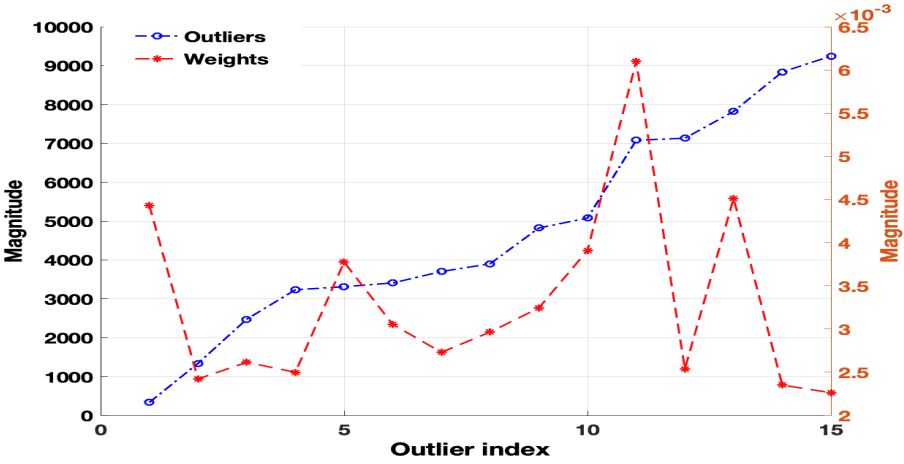
Fig. 10: Water depth prediction values for Case 3: (a) the RODP model predictions without outliers in the training dataset; (b) forecasted curves provided by the RODP and the GP model when there are 25% of outliers in the training dataset; an offset of 0.1 and 0.5 is added to the prediction intervals and error bars, respectively, for the visualization purpose.

the Gaussian distribution $P_{t^*,Li} \sim \mathcal{N}(\mathrm{P}_{t^*,Li}, 0.05\mathrm{P}_{t^*,Li})$ for loads; Weibull distribution for WGs $P_{t^*,G35}, P_{t^*,G36}, P_{t^*,G37} \sim \text{Weibull}(2.06, 7.41)$; and Beta distribution for PVs $P_{t^*,G38} \sim \text{Beta}(2.06, 2.5)$, $i = 1, 2 \ldots, 39$, $t^* = 401, \ldots, 424$.

PV time-series data are often corrupted with spiky signal noise, sensor failure, communication equipment failure, maximum power tracking abnormalities, array shutdown, and power limitation, to name a few (see, Li et al. (2020)), while WG time-series data are frequently corrupted by communication errors, unit outages, and curtailment (see, Ye et al. (2016)). These abnormal conditions are represented by manually added outliers in the measurements, for instance, as shown in Figure 11 for the RES real output powers. To be precise, we add outliers to the measurement vectors $(\mathbf{P}_{G35}, \mathbf{P}_{G36}, \mathbf{P}_{G37}, \mathbf{P}_{G38})$ and to the load consumptions at the load buses, $(\mathbf{P}_{L1}, \mathbf{P}_{L2}, \ldots, \mathbf{P}_{L20})$, up to 25% of the training data points to impose a worst-case scenario that set a benchmark for this study. The probability density estimates of the voltage magnitude and phase angle at the Bus 6 predicted at the instance t_{401} are displayed in Figures 12 (a) and (b), respectively. Their prediction values for the time-interval $[t_{401}, t_{424}]$ are depicted in Figures 12 (c) and (d). As for the nodal voltage magnitude and angle predictions at all the 39 buses at the instance t_{401}, they are depicted in Figures 13 (a) and (b), respectively. The comparison of the results between the conjugate GP regression model and the RODP model obtained for the voltage magnitude at Bus 6 is displayed in Figure 14 along with the RMSE.



(a)



(b)

Fig. 11: Outliers added in the training data: (a) the QQ-plot of the real power measurements with 25% of added outliers and (b) the plot of weights of the SHGM estimator versus magnitude of outliers.

5 CONCLUDING REMARKS

In this paper, we developed a robust regression model based on the SHGM estimator that makes use of the projection statistics to identify and downweight the outliers, including bad leverage points. The RODP model is particularly promising when the training data is contaminated with extreme outliers that lie far away from the data cloud and in the case where the distributions of the errors are unknown. We evaluate the statistical efficiency, influence function, and breakdown point of the RODP model. We test its performance on an artificial numerical two-input model under the Gaussian, Laplace, and Student-t error distributions. We evaluate its performance on two real-world applications, namely, stochastic power flow calculations, and wind speed forecast. Our simulation results showed that the RODP model provides reliable estimations while the conventional GP model largely deviates from the true values when trained on a data set contaminated with only a single outlier. In future

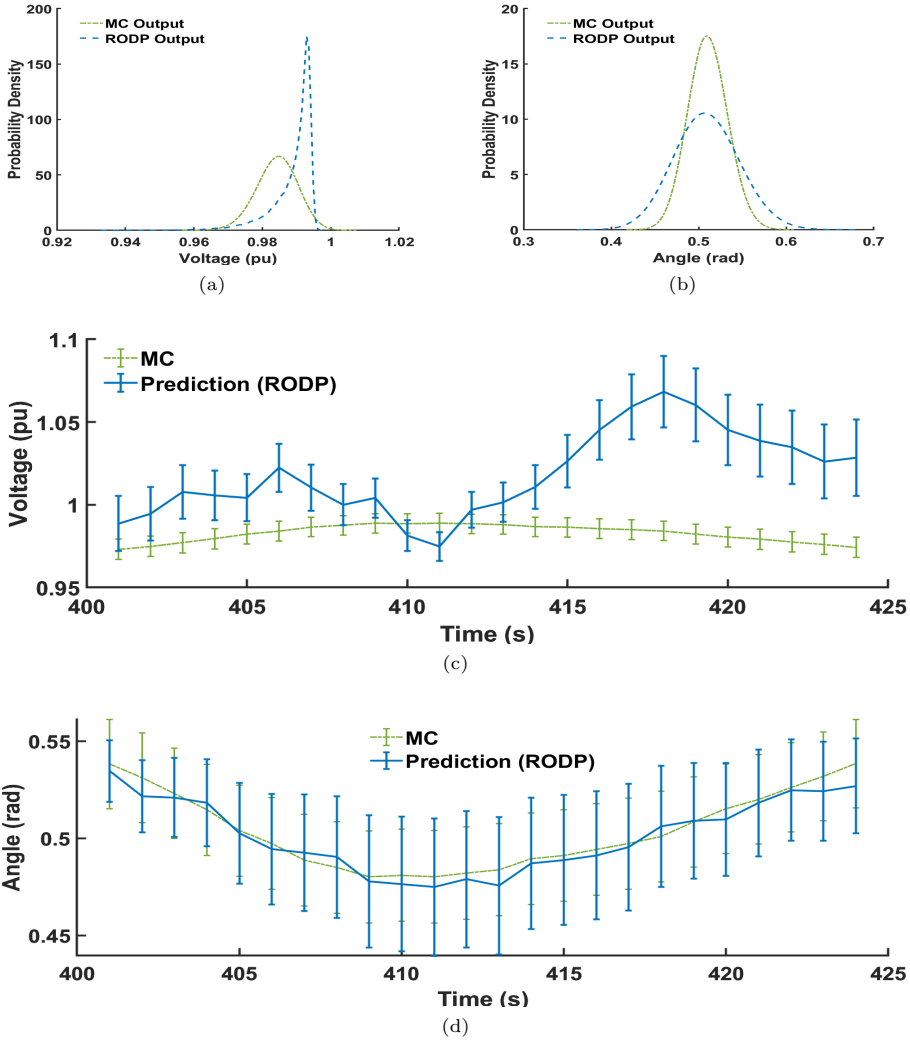
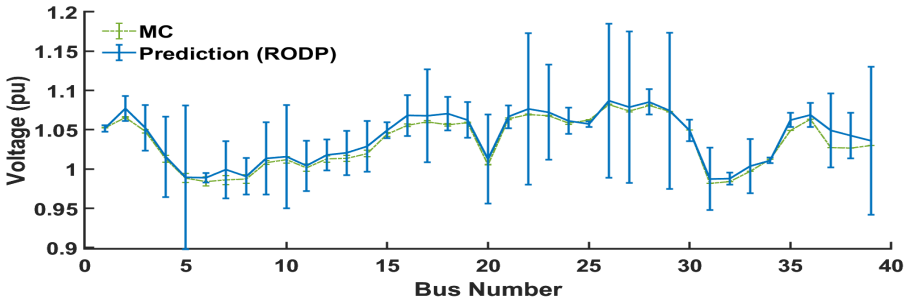
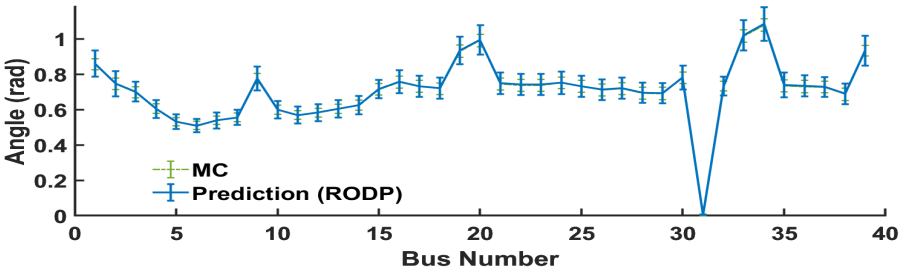


Fig. 12: The RODP model fit through the training data (a) for voltage magnitude and (b) for voltage angle; the prediction values (c) voltage magnitude and (d) voltage angle at Bus 6.

work, we will assess the performance of the RODP model under other thick-tail distributions of error, including the Cauchy distribution and the α stable distributions. We will also apply the developed method to climate modeling (see, Watson-Parris (2021)) and multifractal cosmic signals as suggested by Wirsing et al. (2020). We will also investigate the use of high breakdown estimators to achieve a breakdown point larger than 25% for highly corrupted real-world training datasets.



(a)



(b)

Fig. 13: The RODP model predictions of the IEEE 39-bus power system with the training data corrupted with outliers with magnitudes up to 25% of the original values. Plots of a) nodal voltage magnitudes and b) nodal voltage angles at all buses.

FUNDING

The authors gratefully acknowledge the financial support of NSF via grant ID 1917308.

SUPPLEMENTARY MATERIAL

NGPEMATLAB: Matlab code used for the demonstration of the RODP model on the Gaussian, the Laplace, and the Student-t distributions in Section 4.1. (.tar file)

References

Algikar, P., & Mili, L. (2023), " Robust Gaussian Process Regression with Huber Likelihood.," *arXiv preprint arXiv:2301.07858.* .

Bastos, L. S., & O'Hagan, A. (2009), "Diagnostics for Gaussian Process models," *Technometrics* 51(4), 425–438.

Beaton, A. E., & Tukey, J. W. (1974), "The Fitting of Power Series, Meaning Polynomials, Illustrated on Band-spectroscopic Data," *Technometrics* 16(2), 147–185.

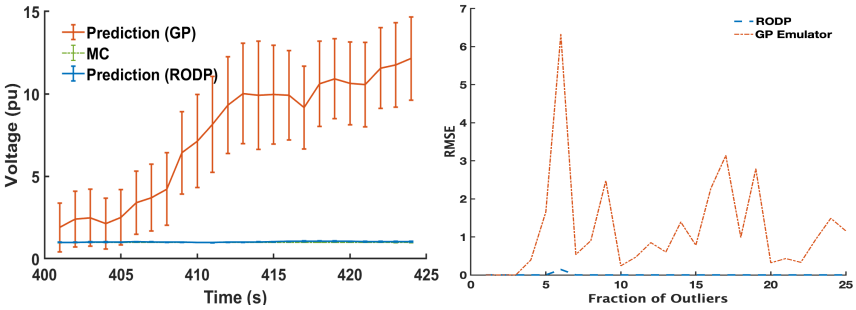


Fig. 14: Comparison between the RODP and GP model at Bus 19 of the IEEE 39-bus power system: (a) voltage magnitude curve at Bus 6 and (b) voltage RMSE curve.

Donoho, D. L. (1982), "Breakdown Properties of Multivariate Location Estimators," *Technical report, Harvard University, Boston,*, URL http://www-stat. stanford. edu/ donoho/Reports/Oldies/BPMLE. pdf.

Donoho, D. L., & Gasko, M. (1992), "Breakdown Properties of Location Estimates Based on Halfspace Depth and Projected Outlyingness," *The Annals of Mathematical Statistics,*, 1803–1827.

Fisher, R. (1925), "Theory of Statistical Estimation," *In Mathematical proceedings of the Cambridge philosophical society (Vol. 22, No. 5, pp. 700-725). Cambridge University Press.*

Gandhi, M. A., & Mili, L. (2010), "Robust Kalman Filter Based on a Generalized Maximum-Likelihood-Type Estimator," *IEEE Transactions on Signal Processing,* 58(5), 2509–2520.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (1986), *Robust Statistics: The Approach Based on Influence Functions, (Vol. 196), John Wiley and Sons.*

Huber, P. J. (1964), "Robust Estimation of a Location Parameter," *The Annals of Mathematical Statistics,* 35(1), 73–101.

Huber, P. J. (2004), *Robust Statistics (Vol. 523),* John Wiley & Sons.

Jylänki, P., Vanhatalo, J., and Vehtari, A. (2011), "Robust Gaussian Process Regression with a Student-t Likelihood," *Journal of Machine Learning Research,* 12(11).

Kuss, M., Pfingsten, T., Csató, L., & Rasmussen, C. E. (2005), "Approximate Inference for Robust Gaussian Process Regression," *Technometrics* 51(4), 425–438.

Li, G., Duan, Z., Liang, L., Zhu, H., Hu, A., Cui, Q., & Hu, W. (2020), "Outlier Data Mining Method Considering the Output Distribution Characteristics for Photovoltaic Arrays and its Application," *Energy Reports* 6, 2345–2357.

Maronna, R. A., & Yohai, V. J. (2016), "The Behavior of the Stahel-Donoho Robust Multivariate Estimator," *Journal of the American Statistical Association,* 429(90), 330–341.

Martinez-Cantin, R., Tee, K., & McCourt, M. (2018, March), "Practical Bayesian optimization in the presence of outliers," *International Conference on Artificial Intelligence and Statistics,* , 1722–1731.

Mili, L., Cheniae, M. G., Vichare, N. S., & Rousseeuw, P. J. (1992), "Robust state estimation based on projection statistics," *IEEE Transactions on Power Systems,* 11(2), 1118–1127.

Oakley, J.E., & O'Hagan, A. (2004), "Probabilistic Sensitivity Analysis of Complex Models: a Bayesian Approach," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 66(3), 751–769.

Murty, P.S.R. (2017), *Power Systems Analysis,* Butterworth-Heinemann , 205–276.

Rasmussen, C. E., and Williams, C. K. (2006), *Gaussian Processes for Machine Learning, (Vol. 2, No. 3, p. 4),* Cambridge, MA: MIT press.

Roberts S., Osborne M., Ebden M., Reece S., Gibson N., & Aigrain S. (2013), "Gaussian Processes for Time-series Modelling," *Philosophical Transactions of the Royal Society* A (371).

Rousseeuw, P. J., & Leroy, A. M. (2005), *Robust Regression and Outlier Detection,* John Wiley & Sons.

Rousseeuw, P. J., and van Zomeren, B. C. (1991), "Robust Distances: Simulations and Cutoff Values," *Directions in Robust Statistics and Diagnostics, Part II* (11), 195–203.

Stahel, W. A. (1981), "Robuste Schätzungen: Infinitesimale Optimalität und Schätzungen von Kovarianzmatrizen," *Ph.D. Thesis, ETH Zürich, Switzerland,.*

Wais, P. (2017), "Two and Three-parameter Weibull distribution in Available Wind Power Analysis," *Renewable Energy, 103*, 15–29.

Wang, X., Liu, R. P., Wang, X., Hou, Y., & Bouffard, F. (2021), "A data-driven uncertainty quantification method for stochastic economic dispatch," *IEEE Transactions on Power Systems, 37(1)*, 812–815.

Wang, X., Wang, X., Sheng, H., & Lin, X. (2020), "A data-driven sparse polynomial chaos expansion method to assess probabilistic total transfer capability for power systems with renewables," *IEEE Transactions on Power Systems, 36(3)*, 2573–2583.

Watson-Parris, D. (2021), "Machine Learning for Weather and Climate are Worlds Apart," *Philosophical Transactions of the Royal Society A, 379*(2194), 20200098.

Wirsing, K., and Mili, L. (2020), "Multifractal Analysis of Geomagnetically Induced Currents Using Wavelet Leaders," *Journal of Applied Geophysics,* 173, 103920.

Xu, Y., Hu, Z., Mili, L., Korkali, M., and Chen, X. (2020), "Probabilistic Power Flow Based on a Gaussian Process model," *IEEE Transactions on Power Systems* 35(4), 3278-3281.

Ye, X., Lu, Z., Qiao, Y., Min, Y., and O'Malley, M. (2016), "Identification and Correction of Outliers in Wind Farm Time-series Power Data," *IEEE Transactions on Power Systems* 31(6), 4197–4205.

Zhang, C., Wei, H., Zhao, X., Liu, T., and Zhang, K. (2016), " A Gaussian Process Regression Based Hybrid Approach for Short-term Wind Speed Prediction," *Energy conversion and management* (126), 1084–1092.

Appendix A Robustness Concepts

Robustness is defined by Huber (1964) as insensitivity to small departures from the assumptions.

- **Influence Function**
  For an ϵ-contaminated model $G = (1-\epsilon)\phi+\epsilon H$, the influence function based on Gateaux derivative that quantifies the local sensitivity of an estimator $T(G)$ to arbitrary infinitesimal contamination $H = \Delta_{\boldsymbol{r}}$ is expressed as

$$\mathrm{IF}(\boldsymbol{r}_i, \boldsymbol{h}_i;\ \phi) = \lim_{\epsilon \to 0} \frac{T((1-\epsilon)\phi + \epsilon\Delta_r) - T(\phi)}{\epsilon}. \tag{A1}$$

  The total influence function (IF), $\mathrm{IF}(r;\ \phi)$, of an M-estimator is equal to the product of the scalar-valued influence of residuals, $\mathrm{IR}((\boldsymbol{r};\ \phi))$, and the vector-valued influence of position, $\mathrm{IP}(\boldsymbol{r};\ \phi)$. Formally, we have

$$\mathrm{IF}(r;\ \phi) = \mathrm{IR}(r;\ \phi)\mathrm{IP}(r;\ \phi). \tag{A2}$$

  They are given by $\mathrm{IR}(\boldsymbol{r};\ F) = \frac{\Psi(\frac{r}{s})}{E\left[\Psi'(\frac{r}{s})\right]}$ and $\mathrm{IP}(\mathbf{f};\ \phi) = (\mathbf{F}^T\mathbf{F})^{-1}\mathbf{f}$. For an M-estimator, the $\mathrm{IR}((\boldsymbol{r};\ \phi))$ is bounded if the $\Psi(\cdot)$ is bounded while $\mathrm{IP}((\boldsymbol{r};\ F))$ is always unbounded, revealing its non-robust to bad leverage points.

- **Asymptotic maximum bias curve**
  The asymptotic maximum bias curve is the curve of the upper bound of a bias of an estimator for an increasing level of contamination $0 \leq \epsilon < \epsilon^*$. The asymptotic maximum bias curve of any Fisher consistent estimator $\hat{\mathbf{x}}$ in its functional form $\boldsymbol{T}$ at any ϵ contaminated model is defined as $b_{max}(\epsilon) = \underset{H}{sup} \mid \boldsymbol{T}(G) - \mathbf{x} \mid$. In the location case, Huber (1964) showed that the sample median has the smallest possible asymptotic maximum bias curve among all location equivariant estimators. In linear regression, the estimator that has the minimum asymptotic bias curve is unknown.
- **Finite-sample breakdown point**
  The maximum value of ϵ, denoted as ϵ^*, for which the maximum bias of an estimator is finite is called the finite-sample breakdown point of that estimator. Formally we have $\epsilon^* = max\{\epsilon;\ b_{max}(\epsilon) < \infty\}$. Donoho et al. (1992) and Huber (1964) showed that the maximum finite-sample breakdown point of any regression equivariant estimator under the assumption of general position is given by $\left\lfloor \frac{m-n}{2} \right\rfloor / m$. Recall that the general position assumption means that any n rows of the observation matrix of a linear regression model are linearly independent.
- **Statistical efficiency** The minimum possible variance that any estimator is able to attend is given by the Cramer-Rao lower bound. Formally, for an

estimator of location ($\hat{\boldsymbol{\theta}}_m$) we have

$$\text{Var}(\sqrt{m}\hat{\boldsymbol{\theta}}_m) \geq \frac{1}{I_f}; \quad \forall m. \tag{A3}$$

An asymptotically efficient estimator attends the Cramer-Rao lower bound with statistical efficiency e_m (see, Fisher (1925)), that is,

$$e_m = \frac{1/I_f}{\text{Var}(\sqrt{m}\hat{\boldsymbol{\theta}}_m)}. \tag{A4}$$

Appendix B Smearing and Masking Effect

Let us consider a linear regression model of the GP model given by

$$\boldsymbol{y} = \mathbf{F}(\mathbf{X})\boldsymbol{\beta} + \boldsymbol{e}, \tag{B5}$$

where $\boldsymbol{e}$ is the error vector whose elements are assumed to be independent random variables following a Gaussian distribution with zero mean and a covariance matrix, $\mathbf{R}$, that is, $\boldsymbol{e} \sim \mathcal{N}(\mathbf{0}, \mathbf{R})$. The residual vector is defined as the difference between the observation vector, $\boldsymbol{y}$, and the estimated vector, $\hat{\boldsymbol{y}}$. Formally, we have

$$\boldsymbol{r} = \boldsymbol{y} - \hat{\boldsymbol{y}}. \tag{B6}$$

For the weighted least squares (WLS) estimator, we have

$$\hat{\boldsymbol{y}} = \mathbf{S}\boldsymbol{y}, \tag{B7}$$

where $\mathbf{S}$ is the hat matrix given by $\mathbf{S} = \mathbf{F}(\mathbf{F}^T\mathbf{R}^{-1}\mathbf{F})\mathbf{F}^T\mathbf{R}^{-1}$. Substituting (B7) into (B6) yields

$$\begin{aligned} \boldsymbol{r} &= \boldsymbol{y} - \mathbf{S}\boldsymbol{y}, & \text{(B8)} \\ &= (\mathbf{I} - \mathbf{S})\boldsymbol{y}, & \text{(B9)} \\ &= \mathbf{W}\boldsymbol{y}. & \text{(B10)} \end{aligned}$$

Substituting (B5) into (54) yields

$$\boldsymbol{r} = \mathbf{W}\boldsymbol{e}. \tag{B11}$$

$\mathbf{W}$ is called the residual sensitivity matrix as it expresses the sensitivity of the residuals to the errors. For the WLS estimator, the outlier detection and identification statistical tests suffer from the smearing and masking effect of the outliers on the residuals as shown next.

- **Smearing effect**
  Let us assume that $e_1 \neq 0$ and $e_i = 0$ for $i = 2, 3, \ldots, m$. From (B11), we have

$$\begin{aligned} r_1 &= W_{11}e_1 \neq 0, \\ r_2 &= W_{21}e_1 \neq 0, \\ r_m &= W_{m1}e_1 \neq 0. \end{aligned}$$

  This is known as the smearing effect of one outlier on the residuals, which makes the identification of that outlier using the residual statistical test difficult to achieve.

- **Masking effect**
  Let us assume that the $e_1 \neq 0$ and $e_2 \neq 0$ while $e_i = 0$ for $i = 3, 4, \ldots, m$. The associated residuals are expressed as

$$r_1 = W_{11}e_1 + W_{12}e_2,$$
$$r_2 = W_{21}e_1 + W_{22}e_2.$$

  Therefore, there exist e_1 and e_2 so that the residuals $r_1 \approx 0$ and $r_2 \approx 0$. This is known as the masking effect of the outliers on the residuals, which results in the failure of the outlier residual identification test.

Appendix C Influence Function of the SHGM Estimator

Theorem: The influence function (IF) of the SHGM estimator over G is given by

$$\text{IF}(r_{S_i}, \boldsymbol{f}_i;\ \Phi) = \frac{\psi(r_{S_i})}{E_\Phi[\psi'(r_{S_i})]}(\mathbf{F}^T\mathbf{F})^{-1}\boldsymbol{f}_i w(\boldsymbol{f}_i), \tag{C12}$$

where r_{S_i} are standardized residuals, $i = 1, 2, \ldots, m$.

Proof The asymptotic influence function of an estimator $\boldsymbol{T}(\cdot)$ is given by

$$\text{IF}(r_i, \boldsymbol{f}_i;\ \Phi) = \lim_{\varepsilon \to 0} \frac{\boldsymbol{T}((1-\varepsilon)\Phi + \varepsilon\Theta_{\boldsymbol{r}}) - \boldsymbol{T}(\Phi)}{\varepsilon}, \tag{C13}$$

$$= \frac{\boldsymbol{T}(G)}{\varepsilon}\bigg|_{\varepsilon=0}. \tag{C14}$$

Let us define the function solved by the SHGM estimator in (17) as

$$\lambda_i(r_i, \boldsymbol{f}_i, \boldsymbol{\beta}) = w(\boldsymbol{f}_i)\boldsymbol{f}_i\psi(r_i). \tag{C15}$$

We express (C15) in the functional form of the SHGM estimator $\boldsymbol{T}(G)$ as

$$\int \lambda(r_i, \boldsymbol{f}_i, \boldsymbol{T}(G))dG = \mathbf{0}. \tag{C16}$$

Let us define the arbitrary infinitesimal contamination at $\boldsymbol{r}$ as $H = \Theta_{\boldsymbol{r}}$. Substituting $G(r) = (1-\varepsilon)\Phi(r) + \varepsilon\Theta_{\boldsymbol{r}}$ into (C16) yields

$$\int \lambda(r_i, \boldsymbol{f}_i, \boldsymbol{T}(G))d\Phi + \int \lambda(r_i, \boldsymbol{f}_i, \boldsymbol{T}(G))d(\Theta_{\boldsymbol{r}} - \Phi) + \varepsilon\frac{\partial}{\partial\varepsilon}\int \lambda(r_i, \boldsymbol{f}_i, \boldsymbol{T}(G))d(\Theta_{\boldsymbol{r}} - \Phi) = \mathbf{0}. \tag{C17}$$

At $\varepsilon = 0$, the last term is zero. Assuming Fisher consistency at Φ that is $\int \lambda(r_i, \boldsymbol{f}_i, \boldsymbol{T}(\Phi)) = \mathbf{0}$, (C17) further reduces to

$$\int \frac{\partial}{\partial\varepsilon}\lambda(r_i, \boldsymbol{f}_i, \boldsymbol{T}(G))\bigg|_{\varepsilon=0} d\Phi + \lambda(r_i, \boldsymbol{f}_i, \boldsymbol{T}(G))d(\Theta_{\boldsymbol{r}}) = \mathbf{0}. \tag{C18}$$

We apply the chain rule to the first term and the shifting property of the Dirac impulse to the second term. After standardizing the residuals to r_{S_i} and solving for the influence function yields

$$\text{IF}(r_i, \boldsymbol{f}_i;\ \Phi) = \frac{\partial \boldsymbol{T}(G)}{\partial\varepsilon}\bigg|_{\varepsilon=0}, \tag{C19}$$

$$= -\left(\int \frac{\partial}{\partial \boldsymbol{\beta}} \lambda(r_{S_i}, \boldsymbol{f}_i, \boldsymbol{\beta}) d\Phi\right)^{-1} \times \lambda(r_{S_i}, \boldsymbol{f}_i, \boldsymbol{T}(G)). \tag{C20}$$

Substituting (C15) in (C19) and expanding the partial derivative yields

$$\mathrm{IF}(r_{S_i}, \boldsymbol{f}_i;\ \Phi) = \left[-\int \boldsymbol{f}_i w(\boldsymbol{f}_i) \frac{\partial \psi(r_{S_i})}{\partial r_{S_i}} \frac{\partial r_{S_i}}{\partial \boldsymbol{\beta}} d\Phi\right]^{-1}_{\boldsymbol{T}(\Phi)} \boldsymbol{f}_i w(\boldsymbol{f}_i) \psi(r_{S_i}), \tag{C21}$$

which reduces to

$$\mathrm{IF}(r_{S_i}, \boldsymbol{f}_i;\ \Phi) = \left[\int \frac{\partial \psi(r_{S_i})}{\partial r_{S_i}} \boldsymbol{f}_i \boldsymbol{f}_i^T d\Phi\right]^{-1} \boldsymbol{f}_i w(\boldsymbol{f}_i) \psi(r_{S_i}) \tag{C22}$$

The integral in (C22) computes the ensemble average of the function of the random vector $\boldsymbol{r}$, and equivalently for the term $\partial \psi(r_{S_i})/\partial r_{S_i}$, for every $(\boldsymbol{f}_i \boldsymbol{f}_i^T)$, $i = 1, 2, \ldots, m$. Therefore, we factor out the mean of $\partial \psi(r_{S_i})/\partial r_{S_i}$ that yields

$$\int \frac{\partial \psi(r_{Si})}{\partial r_{S_i}} \boldsymbol{f}_i \boldsymbol{f}_i^T d\Phi = E_\Phi[\psi^{'}(r_{S_i})](\mathbf{F}^T \mathbf{F}). \tag{C23}$$

Substituting (C23) into (C22) gives the following expression of the influence function

$$\mathrm{IF}(r_{S_i}, \boldsymbol{f}_i;\ \phi) = \frac{\Psi(r_{S_i})}{E_\phi[\Psi'(r_{S_i})]} (\mathbf{F}^T \mathbf{F})^{-1} \boldsymbol{f}_i w(\boldsymbol{f}_i). \tag{C24}$$

□

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- NGPEMATLAB.zip